

# *Distributed Data Mining Techniques for Object Discovery in the National Virtual Observatory (NVO)*

**Kirk Borne, George Mason University, GSFC**  
**Cynthia Cheung, NASA, GSFC**

## References

<http://nvo.gsfc.nasa.gov/>

e-mail: [Kirk.Borne@gsfc.nasa.gov](mailto:Kirk.Borne@gsfc.nasa.gov)

# Abstract

We are pursuing an exploratory data mining project to identify classification features of special classes of interacting galaxies (for example, infrared-luminous galaxies) within distributed astronomical databases. Using a variety of data mining techniques, interaction-specific features are learned -- in order to distinguish this class of galaxies from a control sample of normal galaxies. Subsequently, the corresponding rule-based feature model of that class of galaxies is then applied to the large multi-wavelength astronomical databases that are now becoming available. This distributed data mining activity is a prototype science use case for the NVO (National Virtual Observatory). We specifically apply multi-archive multi-wavelength data to the problem. We are researching both successful and unsuccessful data mining attempts.

# What is Data Mining? . . .

***Data Mining is an information extraction activity whose goal is to discover hidden facts contained in large databases.***

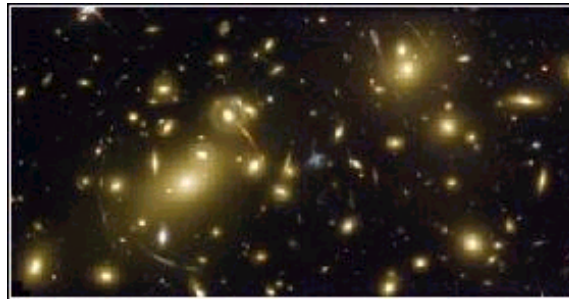


- The end goal of data mining is not the data themselves, but the new knowledge and understanding that are revealed in the process (i.e., KDD = Knowledge Discovery in Databases) :

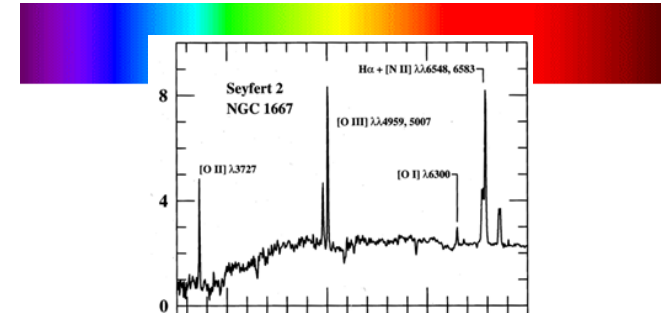
**Data → Information → Knowledge → Understanding**

# Astronomy Example

## Data:



(a) Imaging data (ones & zeroes)



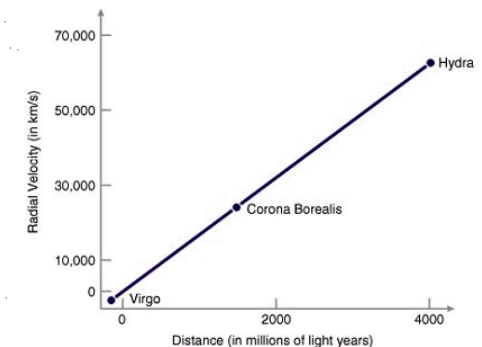
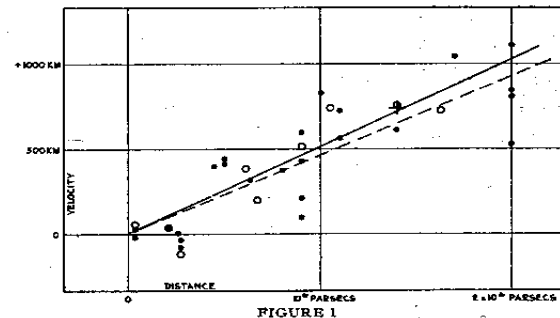
(b) Spectral data (ones & zeroes)

## Information (catalogs / databases):

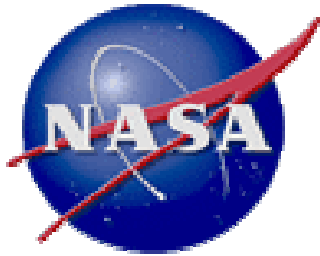
- Measure brightness of galaxies from image (e.g., 14.2 or 21.7)
- Measure redshift of galaxies from spectrum (e.g., 0.0167 or 0.346)

## Knowledge:

Hubble Diagram →  
Redshift-Brightness  
Correlation →  
Redshift = Distance



**Understanding:** the Universe is expanding!!



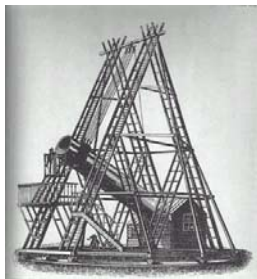
Searching, Retrieving, Mining, Integrating, and Analyzing geographically distributed Astronomical Data Repositories is one of the key goals of VO (Virtual Observatory) projects:

**The VO is "Distributed Data Mining in Action"**



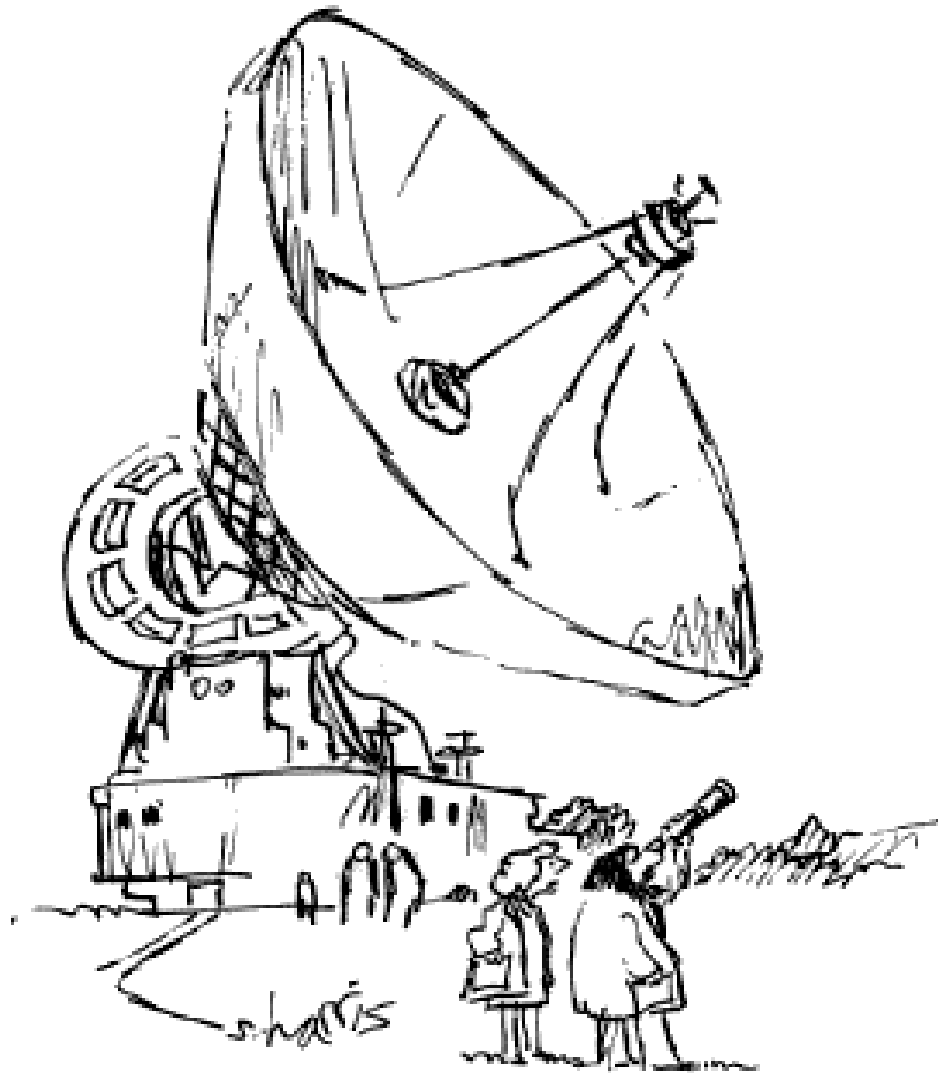
# The National Virtual Observatory

- National Academy of Sciences “Decadal Survey” recommended NVO as highest priority small (<\$100M) project :  
*“Several small initiatives recommended by the committee span both ground and space. The first among them—the National Virtual Observatory (NVO)—is the committee’s top priority among the small initiatives. The NVO will provide a “virtual sky” based on the enormous data sets being created now and the even larger ones proposed for the future. It will enable a new mode of research for professional astronomers and will provide to the public an unparalleled opportunity for education and discovery.” (p.14)*



Astronomy and Astrophysics  
in the  
New Millennium

# Why so many astronomical databases? ...



"Just checking."

**Because ...**

**Many great astronomical discoveries have come from inter-comparisons of various wavelengths:**

- **Quasars**
- **Gamma-ray bursts**
- **Ultraluminous IR galaxies**
- **X-ray black-hole binaries**
- **Radio galaxies**
- ...

# National Virtual Observatory (NVO)

<http://www.us-vo.org/>

- ▶ NVO is **a concept**, recommended by the National Academy of Sciences.
- ▶ NVO will link geographically distributed astronomical data archives and information resources = **provides “one-stop shopping” for data end-user.**
- ▶ NVO will be heterogeneous, **interoperable**, and federated (autonomy maintained at local sites) ... therefore, **XML** and Web Services.
- ▶ NVO requires **innovative information technologies** for :
  - ▶ **data discovery, data mining, data fusion, distributed querying.**
- ▶ Volume of archived data for an all-sky survey :
  - One band = 4 Terabytes
  - Multi-wavelength = 10-100 Terabytes
  - Time dimension = 10 Petabytes
  - LSST project (10 yrs) = ~100 Petabytes <<http://www.lsst.org/>>



# Why is it necessary?

- ▶ To maximize cross-enterprise multi-institutional resources
- ▶ To minimize duplication of effort
- ▶ To streamline operations through shared development
- ▶ To serve multiple user communities
- ▶ To facilitate simultaneous data mining, knowledge discovery, and information retrieval from multiple distributed data collections
- ▶ Because data volumes are huge & growing rapidly ...

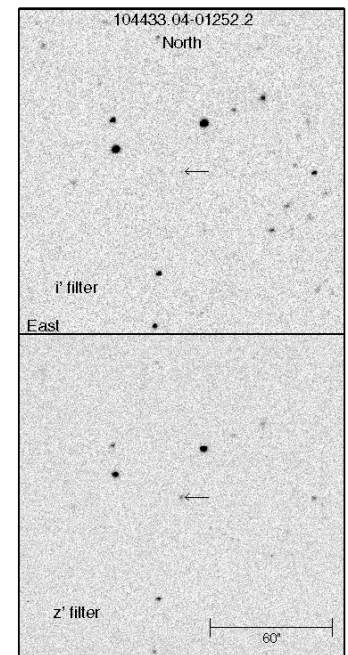
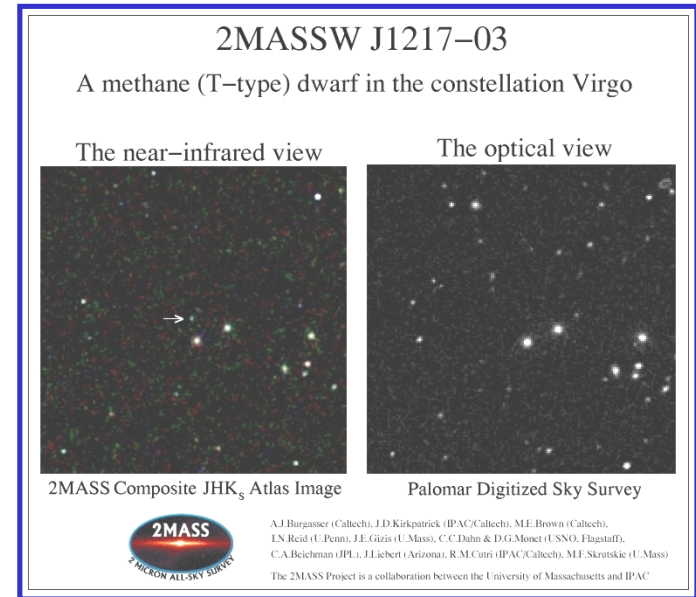
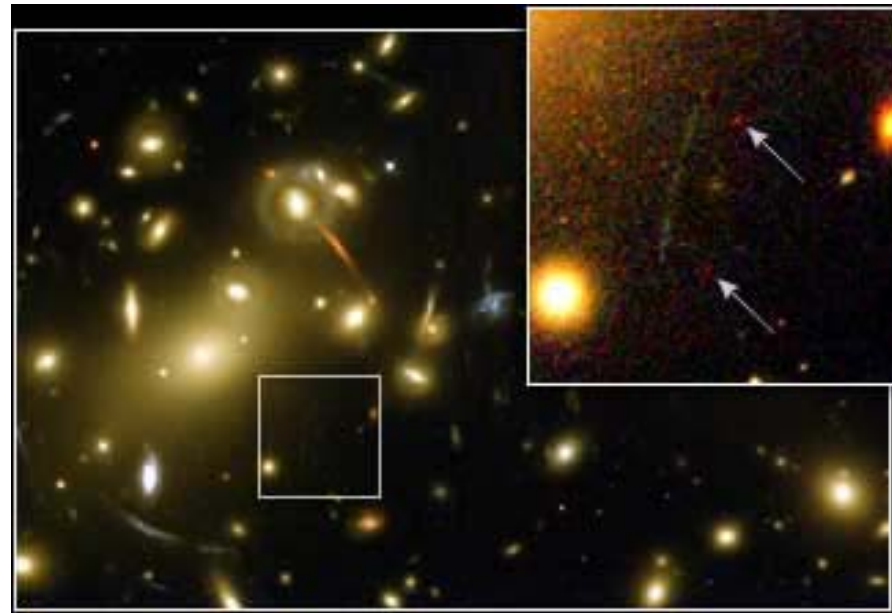
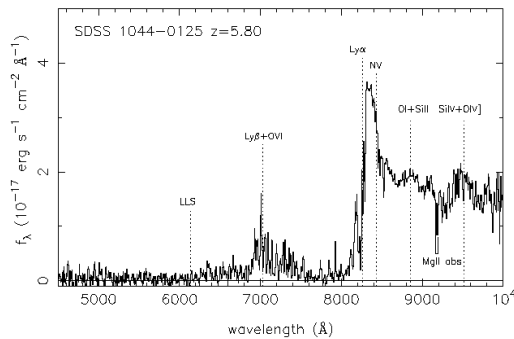
For example, in Astronomy :

- ▶ a few terabytes "yesterday" (10,000 CDROMs)
- ▶ tens of terabytes "today" (100,000 CDROMs)
- ▶ petabytes "tomorrow" (within 10 years) (100,000,000 CDROMs)

# Main reason ... NVO enables new science


<http://www.us-vo.org/>

- Rare and exotic objects
  - Very high redshift quasars
  - Dark matter in the galactic halo
  - Time-variable objects, transient events: distant supernovae and microlensing
  - Brown dwarfs
  - Variable stars
  - Asteroids...
    - ...incoming!!



# NVO Science Cases & Drivers

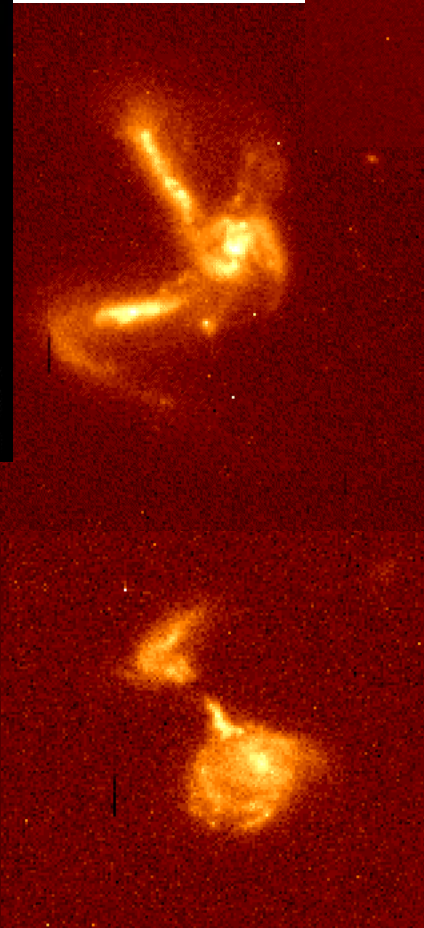
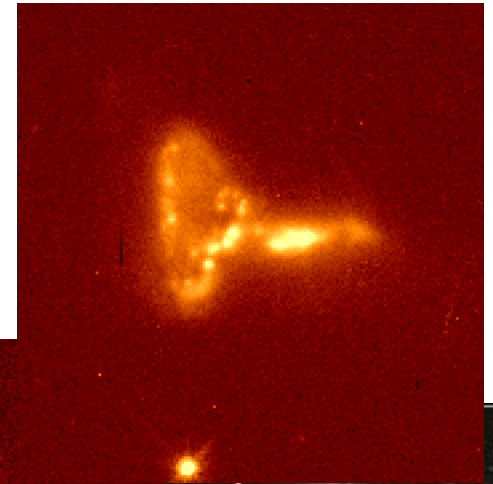
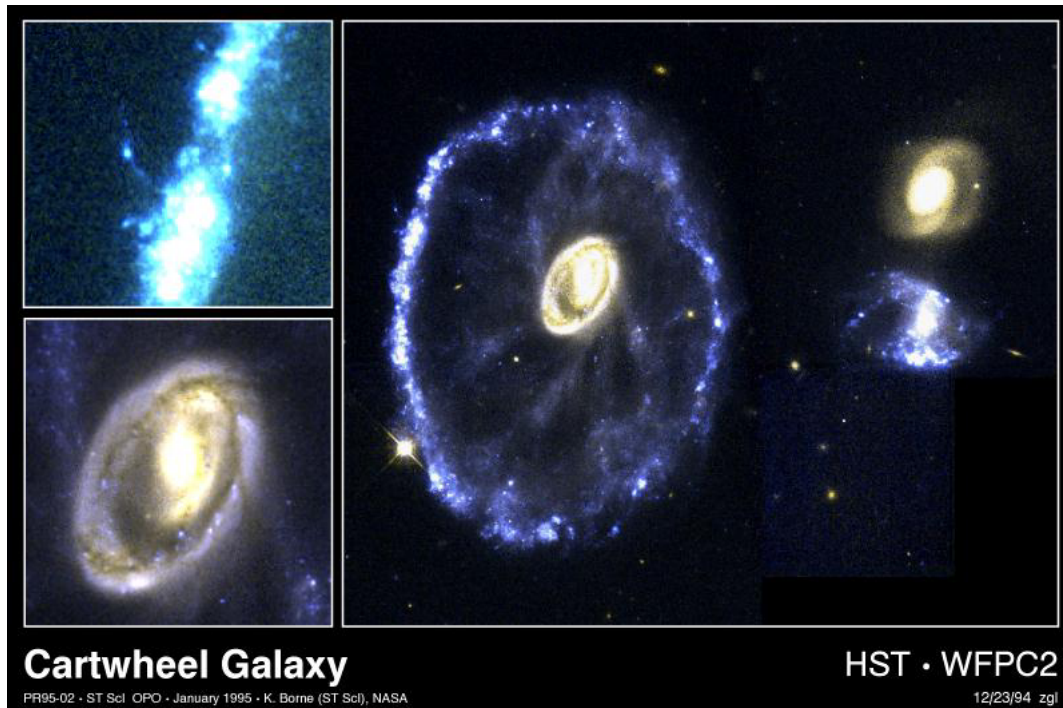
## (from Aspen 2001 NVO Workshop)

- ▶ **Solar System** : NEOs, Long-Period Comets, TNOs, **Killer Asteroids!!!**
  - ▶ **The Digital Galaxy** : Find star streams and populations -- relics of past/present assembly phase. Identify components of disk, thick disk, bulge, halo, arms, ??
  - ▶ **The Low-Surface Brightness Universe** : spatial filtering, multi-wavelength searches, intersection of the image and catalog domains
  - ▶ **Panchromatic Census** of AGN (Active Galactic Nuclei) : Complete sample of the AGN zoo, their emission mechanisms, and their environments
  - ▶ **Precision Cosmology** & Large-Scale Structure : **\*\*Hierarchical Assembly History of Galaxies and Structure\*\***, Cosmological Parameters, Dark Matter and Galaxy Biasing as  $f(z)$
  - ▶ **Precision science of any kind** that depends on very large sample sizes
  - ▶ "Survey Science Deluxe"
  - ▶ **Search for rare and exotic objects** (e.g., high- $z$  QSOs, high- $z$  SNe, L/T dwarfs)
  - ▶ **Serendipity** : Explore new domains of parameter space (e.g., time domain, or "color-color space" of all kinds)
- 

**\*\*This is the scientific goal of the IDU-funded project described here.**



# Colliding and Merging Galaxies: Building Blocks of the Universe

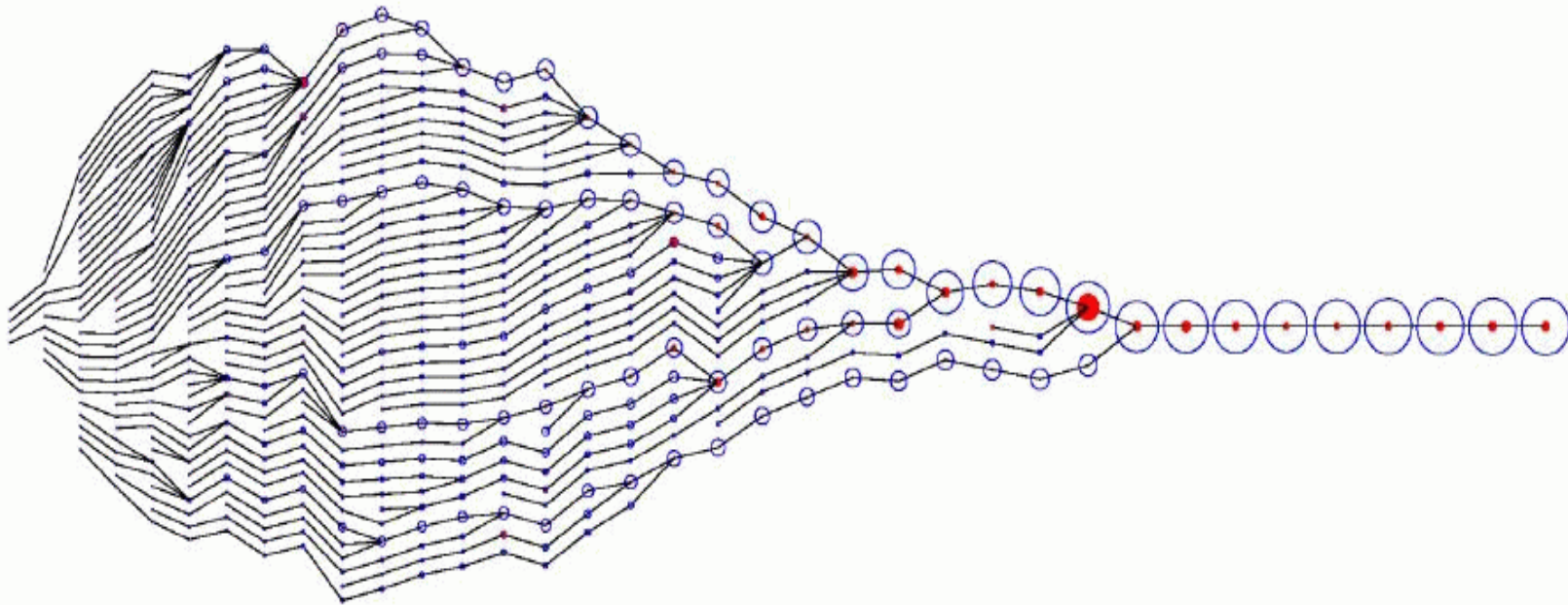


# Ultra-Luminous Infrared Galaxies (ULIRGs) and other IR-Luminous Galaxies (LIRGs)



- Nearly 100% are involved in collisions and mergers.
- LIRGs are among the most luminous galaxies in the Universe.
- Related to various phenomena: Quasars, GRBs, IR background, EROs, super-starbursts, Scuba submm sources.
- May be the missing link in the evolution of young galaxies to Quasars to today's galaxies.
- Provide keys to: (1) Galaxy Assembly (from the cosmic soup following the big bang). (2) Star Formation. (3) Metal Production. (4) Galaxy Evolution. And (5) Massive Black Hole Formation.

# Merger Tree - Galaxy Merger Family History



**Past**

**Present**

The goal of this study is to identify collision and merger remnant candidates at increasing redshift, in order to measure the galaxy hierarchical mass assembly rate as a function of cosmic epoch.



# Distributed Data Mining in the NVO: A case study to find colliding, interacting, and merging galaxies among the IR-luminous galaxies

- NASA IDU Program: (why we are here today)
  - Distributed Data Mining in the NVO
  - Examine several distributed databases (HST, 2MASS, Sloan, IRAS)
  - Solve a particular science problem (a NVO science scenario)



NASA  
Information  
Power Grid  
(IPG)

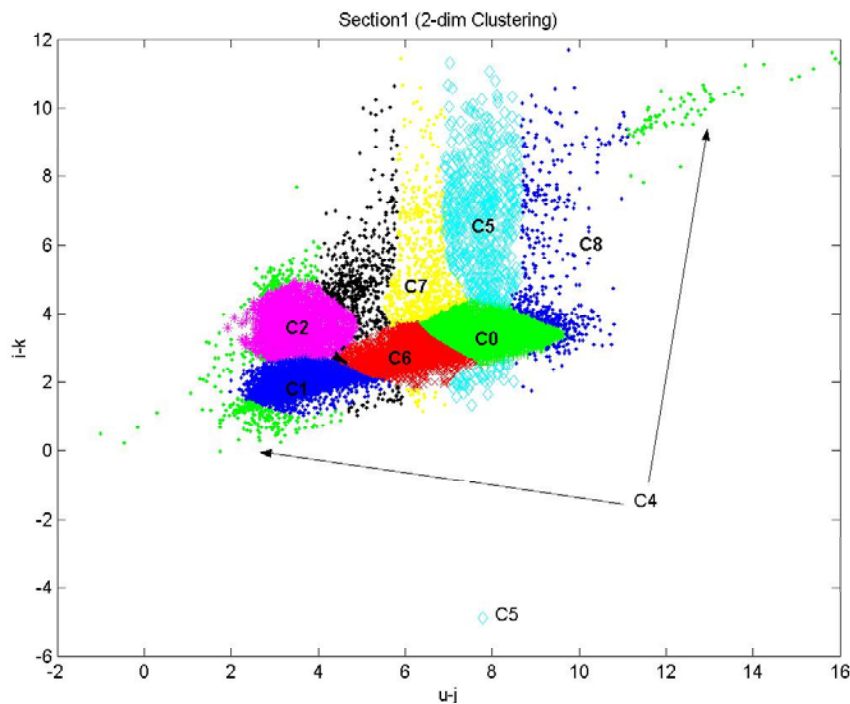
# Distributed Data Mining in the NVO

- NASA IDU Program:
  - K.Borne (PI), Cynthia Cheung (Collaborator), GMU student (scientific programmer), and Hillol Kargupta (UMBC consultant, P.I. on IDU-funded Distributed Data Mining project).
  - 1. Identify all known examples of ULIRGs:
    - linked to Gamma-Ray Bursts, Quasars, Hierarchical Galaxy Assembly, etc.
  - 2. Learn new properties of ULIRGs (e.g., Association Rule Mining) by examining multiple distributed databases.
  - 3. Build a classifier from these rules.
  - 4. Find new cases of ULIRGs in the distributed databases.
  - 5. Results will contribute to understanding of many classes of astronomical phenomena, including JWST science program.
  - 6. Techniques will be applicable to NVO, LWS, other VOs,...

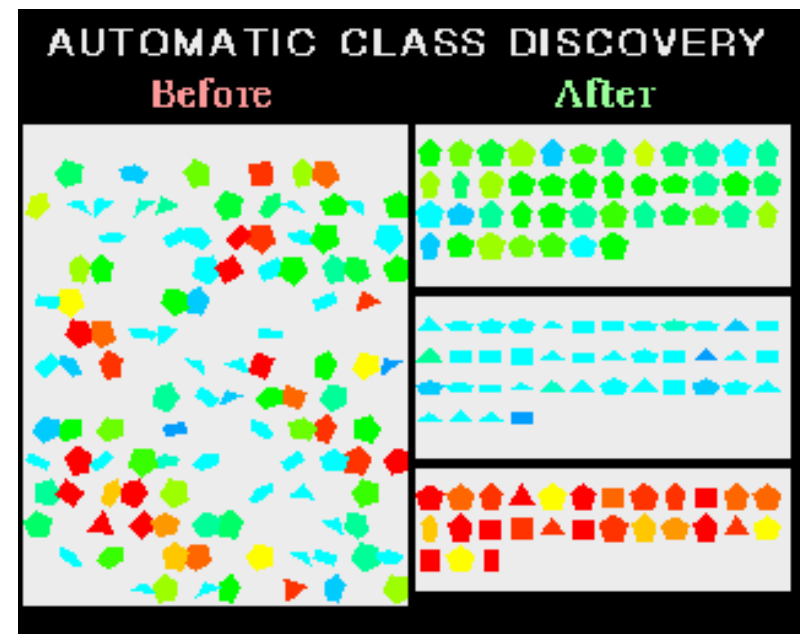


# Class Discovery through Clustering:

Use data mining methods to identify classes of galaxies among several large photometric catalogs (e.g., 2MASS, Sloan, NVSS, etc. --without using redshifts): the galaxy class is either **normal** or **IR-luminous** (indicative of collision/merger activity)



Plot provided by H.Kargupta (UMBC)



(sample data only are displayed in these plots)

An example of clustering in a 3-dimensional color-color parameter space using data from two different (distributed) astronomical databases. In this case, the 3 colors are pairings of 2MASS near-IR and Sloan optical magnitudes.



## Additional application areas of IDU-funded NVO data mining project

- Application of XML to distributed data mining:
  - investigate VOQL (V.O. query language)
  - investigate XMLA (XML for Analysis)
  - investigate PMML (Predictive Modeling Markup Language)
- Application of different data mining techniques:
  - Bayes classification
  - Neural nets
  - Decision trees
  - Association rule mining
  - Genetic Algorithms for rapid data modeling
  - Supervised and Unsupervised Learning algorithms for robust classification
- Application of Beowulfs to parallel high-performance data mining
- Application to new mission data sets: GALEX, Spitzer, WISE, JWST, LWS, Sensor Webs, Constellations (distributed Sciencecraft)

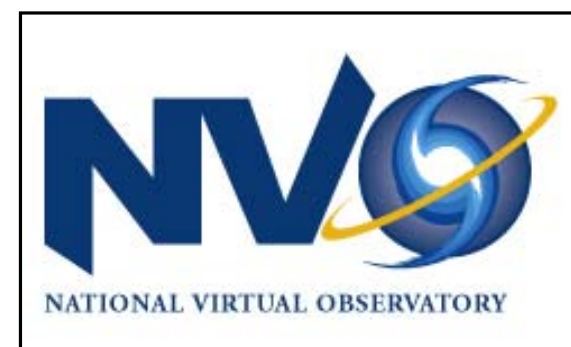
# Summary - Applications of Data Mining to the NVO

## Data Mining Resource Guide for Space Science:

[http://nvo.gsfc.nasa.gov/nvo\\_datamining.html](http://nvo.gsfc.nasa.gov/nvo_datamining.html)

### Sample Data Mining Applications within the NVO:

- Discover data stored in geographically distributed heterogeneous systems.
- Search huge databases for trends and correlations in high-dimensional parameter spaces: identify new properties or new classes of objects.
- Search for rare, one-of-a-kind, and exotic objects in huge databases.
- Identify temporal variations in objects from millions or billions of observations.
- Identify moving objects in huge survey catalogs and image databases.
- Identify parameter glitches / anomalies / deviations either in static databases (e.g., archives) or in dynamic data (e.g., science / telemetry / engineering data streams from remote satellites).
- Find clusters, nearest neighbors, outliers, and/or zones of avoidance in the distribution of astrophysical objects or other observables in arbitrary parameter spaces.
- Serendipitously explore the huge databases that will be part of the NVO, through access to distributed, autonomous, federated, heterogeneous, multi-wavelength, multi-mission astrophysics data archives.



<http://www.us-vo.org/>